The Behavior Analyst 1999, 22, 99–103 No. 2 (Fall)

# Statistical Inference in Behavior Analysis: Having My Cake and Eating It?

Michael Davison
Auckland University

Using simple, nonparametric statistical procedures can formalize the process of letting data speak for themselves, and can eliminate the gratuitous dismissal of deviant data from subjects or conditions. These procedures can act as useful discriminative stimuli, both for behavior analysts and for those from other areas of psychology who occasionally sample our journals. I also argue that changes in publication policies must change if behavior analysts are to accurately discriminate between real, reliable effects (hits) and false alarms.

*Key words:* nonparametric statistics, Type I error, Type II error, conservatism

When I was invited to take part in the panel discussion that was the basis for this paper, I began to worry about my behavior in relation to statistics. I guess that one of the reasons that I was asked to take part was that I habitually use inferential statistics in my papers. Given that, perhaps, the unanalyzed life is not worth living, I started wondering why I did this. After searching my behavioral soul, my world-shattering conclusion is that my behavior is the product of my history. Let me lay bare some of this to you.

While I was working on my doctorate in Dunedin, New Zealand, my supervisor and I conducted some research concerned with what controlled the pause after reinforcement in fixed-ratio schedules. The data were exceedingly clear: The next ratio requirement was the major variable, though there were some effects of prior requirements. My supervisor suggested that I take these results to a conference in Australia, and I naturally agreed. "Oh, by the way, Australians are very keen on statistics, so you'd better do an analysis of variance," he said. So I did, being very keen to impress potential employers. I gave the paper, showed the data in all their glory, and then presented an ANOVA. Question time, and one person got up and said "That was the wrong ANOVA model, you should have used this one." A second person said the first person was wrong, it should have been that one. And then all hell let loose. There were no bouquets, questions about the data, the experimental procedure, nothing. There were no reinforcers. As we all know, one bad experience can sully your whole life, and this did. Until quite recently, I never did another parametric analysis of variance. I generalized from this to all parametric statistics, and have shied away almost completely from these henceforth.

## *Individual Differences*

Over the many years that I have been writing for the *Journal of the Experimental Analysis of Behavior,* reviewers have vacillated from requiring no statistics, through requiring variance measures in data (but no inferential statistics), to requiring some sort of statistical treatment. From this extended training, and from my reading in the area, I came to the conclusion that not using statistics was a bad practice, and that at least some formalization of the process that we use to determine the meaning of data was required. One of the processes that upset me the most is the dismissal of results from single subjects or single condi-

I thank Susan Schneider and Douglas Elliffe for useful and informed comments on a draft of this paper.

Reprints may be obtained from Michael Davison, Department of Psychology, Auckland University, Private Bag 92019, Auckland, New Zealand (E-mail: m.davison@auckland.ac.nz).

tions. There is a great variation between researchers in the skill of arguing away deviant data, and I also suspect that a person's standing in the field can have a tremendous influence on whether such arguments are accepted in the editorial process or not. This last influence is perhaps reasonable, but for none, even the most respected, should it be allowed to wag the dog.

I understand the argument for dismissing deviant data. It is entirely possible that some subpopulation of my sample might, because of a different behavioral history, respond quite differently to an experimental manipulation. I probably would want to argue that this is a quantitative difference, rather than a qualitative difference. Within the parameters of a single experiment that did not investigate this quantitative difference, however, it would look qualitative. If this difference really exists, the behavior of one of my subjects might be different. But it would be hard to discriminate between "my finding is not general" and "I had one odd bird" because these are the selfsame conclusions. What I cannot conclude, however, is the all-too-common conclusion that the finding *is* general and that I had an odd bird. This conclusion is tantamount to making the assumption that the odd behavior comes from random error variance, rather than from systematic variation in an independent variable that has not been investigated. Knowing that behavior is multiply caused, I should conduct a follow-up experiment to determine the reason for the odd behavior. Arguing away the deviant subject, though, is functionally equivalent to using an increased $N$ in an inferential statistical model. And as most behavior analysts recognize, inferential statistics gloss over deviant behavior and, depending on how large $N$ is, will either accept or reject the null hypothesis (Hopkins, Cole, & Mason, 1998).

*Sample Size*

Just as dismissing deviant subject performance is not good practice, increasing the sample size in order to minimize individual differences is also not good practice. Sample size does matter, however. If it is the case that $N = 1$ is enough for a radical behaviorist, then, I am not (if this is a requirement for membership), and do not want to be, a radical behaviorist. I am interested in the external validity of my findings, and I feel the need to replicate across subjects—I need to have some idea that the results from a single subject were not an isolated, odd occurrence. But if the membership criteria include being able to show the same effect with all the subjects that we use, I will join. But the question is what $N$ is enough? From my view, 3 or 4 is not enough, *especially* in those cases in which one of the subjects did something different, and its data are argued away. As you will see from my publications, I feel that I need 6 subjects to be able to get a good idea of what is going on. Given that most of my experiments are quite long, I can lose 1 subject and still feel happy. But, because I am interested in the behavior of individual organisms, using 6 subjects rather than 3 or 4 will increase the likelihood of my sampling the subject that behaves differently.

*Nonparametric Statistics*

I generally use nonparametric statistics when reporting my research. Such statistical procedures are appropriate for small numbers of subjects, but I cannot say, now, which is the chicken and which is the egg: Do I use 6 subjects because that gives me a result on a sign test if *all* subjects behave in the same way? Or do I use nonparametric statistics (rather than no statistics at all) because I generally have 6 subjects? I suspect that this is just a dynamic system that has come to feed on itself. However, I do believe that the use of such statistical procedures levels the playing field between researchers, and it does represent rather nicely the much more informal process of letting the data speak for themselves. Or would

do so, if most researchers used a similar number of subjects. I simply believe that 6 out of 6 (even 5 out of 5) is enough to give everyone the confidence that the findings have decent generality. I'd like to see this as a recommendation, rather than a rule.

Simple nonparametric statistics, I think, simulate the best behavior of researchers when they look at their data, and formalize the process of assessment. I think they can act as good discriminative stimuli for a reader's interpretation of the data, although there are dangers in doing so. If I simply report that a slope measure increased between two conditions for all 6 subjects, does this have the same impact as the additional test (significant on a sign test at $p < .05$)? In a sense it should, but for many readers, particularly those from other areas who occasionally sample behavioral journals, it is not. Behavior analysis needs such people to be impressed with its research and cannot afford to have them discard the work with the thought "I bet that isn't significant." Beyond statistics as a mathematical procedure, they are discriminative stimuli for behavior that because of the training of psychologists, have a substantial impact. Some behavior analysts may find this to be unfortunate, but if behavior analysis is to survive it has to live within a larger verbal community whose behavior is affected by the use of statistical tests.

One of the dangers of using statistics, however, is using them incorrectly. If you do the wrong test (like using the wrong ANOVA model!) you can get significance where there is none, and none where there is. Thus, especially for outside readers, behavior analysts also need to make sure they know what they are doing when they use statistics, and to report all germane parameters of the test (and the data used) so this can be checked. In this way behavior analysts end up being as much statisticians—perhaps even more so, given our bad press on this matter—than other psychologists.

| | Effect found | No effect found |
|---|---|---|
| There is an effect | 2 - P | |
| There is no effect | | |

| | Effect found | No effect found |
|---|---|---|
| There is an effect | | |
| There is no effect | 2 - P | |

Figure 1. Two published (P) findings of a "reliable" effect.

### *Conservatism*

These considerations bring me to a further point: conservatism. Again, if joining the radical behavioral club requires me to be conservative in reporting differences, then I want in. Sidman (1960) was right when he insisted that our business is to look for generalities and invariances (Nevin, 1984) rather than differences. It seems to me that this is the real business of science, and that the statistical model (and even the nonstatistical, "let the data speak for themselves" model) can lead, through the interaction with publication processes, to psychology becoming a Type I error (Davison, 1998). The culture of psychology writ large produces students whose main aim in life is to find a significant difference (and hence, publish). We really do have to change this culture, and I believe it is much less prevalent in the experimental analysis of behavior than in the remainder of psychology.

As I have argued before, statistical analyses (or small *N* visual analyses) in combination with publication policies can have a disastrous effect on what is "known" (Davison, 1998). A signal-detection analysis, as shown in Figures 1 and 2, makes this clear. Take two situations, with an effect being reported in both. In Situation A (Figure 1), there is a real effect, and there are two published reports of this effect. The published data suggest a high

| | Effect found | No effect found |
|---|---|---|
| There is an effect | 2 - P | 1 - U |
| There is no effect | 0 | 0 |

| | Effect found | No effect found |
|---|---|---|
| There is an effect | 0 | 0 |
| There is no effect | 2 - P | 8 - U |

Figure 2. Reality. Two published (P) findings of a "reliable" effect, and some unpublished (U).

probability of a "hit," the reporting of a signal when one exists. In Situation B, there is no real effect, but there are also two published findings of an effect—two false alarms. Situation B, of course, could happen by chance. Subsequently, in both cases, no-effect results (misses in Situation A and correct rejections in Situation B) will be very hard to publish, and the journals report in both situations a 2:0 score line for "effect found." We cannot discriminate from the published data which effect is real. For both situations, the initial finding and its replication suggest an effect, but the real situation may be quite different. If misses in Situation A and correct rejections in Situation B were published, however, we might have more data to make the discrimination. For example, if there is one miss in Situation A (Figure 2), there is a reasonably certain decision ($p = .67$) that the effect exists. If in Situation B, however, there are eight correct rejections, then the probability that the effect exists is .2. Thus, the "reality" afforded by publication is strongly biased towards an effect being accepted regardless of whether it is real or not. It follows then, that unless we publish good-quality failures to replicate, we are systematically blinding ourselves to reality, and we cannot discriminate reality from fantasy.

It is also for reasons of conservatism, and the quest for generality, that I use nonparametric statistics. I would rather assume a lower level of measurement in my data than a higher level that my data may not reach. I would rather not assume that my data are normally distributed, because with $N = 6$, I cannot ever demonstrate that they are. I am aware that if the data were normally distributed, I could use more powerful tests, but I would rather deal with exact probabilities rather than approximations under a string of assumptions. I was strongly influenced by the first two chapters of Bradley (1968), which gives a very robust comparison of parametric and nonparametric tests. Finally, I suspect that our high levels of experimental control lead inevitably to the nonnormality of data distributions.

A wealth of nonparametric statistical tests are readily available. They range from binary comparisons, through analysis of variance with post hoc testing and orthogonal polynomial analysis, to regression, for all levels of measurement. Some of the tests, like rank randomization and normal-scores tests, are extremely powerful, with asymptotic relative efficiencies greater than 100% for nonnormal data in comparison with classical $t$ and $F$ tests. My particularly valued resources are Conover (1980), Ferguson (1965), and Marascuillo and McSweeney (1977), which provide an excellent coverage of useful tests. Although I recommend that readers look at other nonparametric tests, I try to keep my usage to simple sign and binomial tests, nonparametric trend tests, and Friedman analysis of variance. These seem to me to be honest and understandable and to fit closely with my own assessment of the importance of effects. Of course, unless I am well out of date, nonparametric tests do not offer analyses of interactions—for which I am eternally grateful. I have often argued, and will continue to do so, that significant interactions indicate wrong measures—yes, they can often be eliminated by transformations, but the goal of basic science must be to discover measures

that are independent of one another. Just think of the havoc that would be created if, in Ohm's law ($V = IR$), $I$ and $R$ had an interactive effect on $V$. (Actually, in any real system, this is exactly what happens as the amps that flow will warm the resistor and change its resistance, but I am talking fundamental, rather than applied, science here.)

Before concluding, though, with an exhortation that nonparametric statistics in our research should be used to provide a conservative playing field, there is one more problem that may arise. The subject matter of psychology is dynamically interacting behavior–environment systems (Davison, 1998). This may make experimental design difficult (independent variables are dependent, and dependent variables are independent, in some sense), and may make data analysis more difficult. In most of psychology and the experimental analysis of behavior, independent variables are nominal, or maybe ordinal. Either way, in any real system, they will have variance, and the arranged nominality or ordinality of the independent variables across subjects may be seriously compromised. It thus becomes important, even essential, to carry out analyses using procedures, such as linear regression (which are descriptive, rather than inferential) that take into account the continuous quantitative nature of the environmental input and its variance. Such regression procedures (known as structural relations and nonparametric regression procedures) are not widely used (see Davison & McCarthy, 1981, for an example) and are in need of further development and, especially, advertisement. It seems to me that assuming no variance in our "independent" variables is an error that is right up there with the global assumption that data are distributed normally (or that it matters not if they aren't)—nicely termed the "normal mystique" by Bradley (1968).

## REFERENCES


Bradley, J. V. (1968). *Distribution-free statistical tests.* Englewood Cliffs, NJ: Prentice Hall.

Conover, W. J. (1980). *Practical nonparametric statistics* (2nd ed.). New York: Wiley.

Davison, M. (1998). Experimental design: Problems in understanding the dynamical behavior-environment system. *The Behavior Analyst, 21,* 219–240.

Davison, M., & McCarthy, D. (1981). Undermatching and structural relations. *Behaviour Analysis Letters, 1,* 67–72.

Ferguson, G. A. (1965). *Nonparametric trend analysis.* Montreal, Canada: McGill University Press.

Hopkins, B. L., Cole, B. L., & Mason, T. L. (1998). A critique of the usefulness of inferential statistics in applied behavior analysis. *The Behavior Analyst, 21,* 125–137.

Marascuillo, L. A., & McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences.* Monterey, CA: Brooks/Cole.

Nevin, J. A. (1984). Quantitative analysis. *Journal of the Experimental Analysis of Behavior, 42,* 421–434.

Sidman, M. (1960). *Tactics of scientific research.* New York: Basic Books.